GROUP 4

Lisbon Airbnb Analysis

Data Pipeline Project

Meet Our Team

KEREN VASCONCELOS

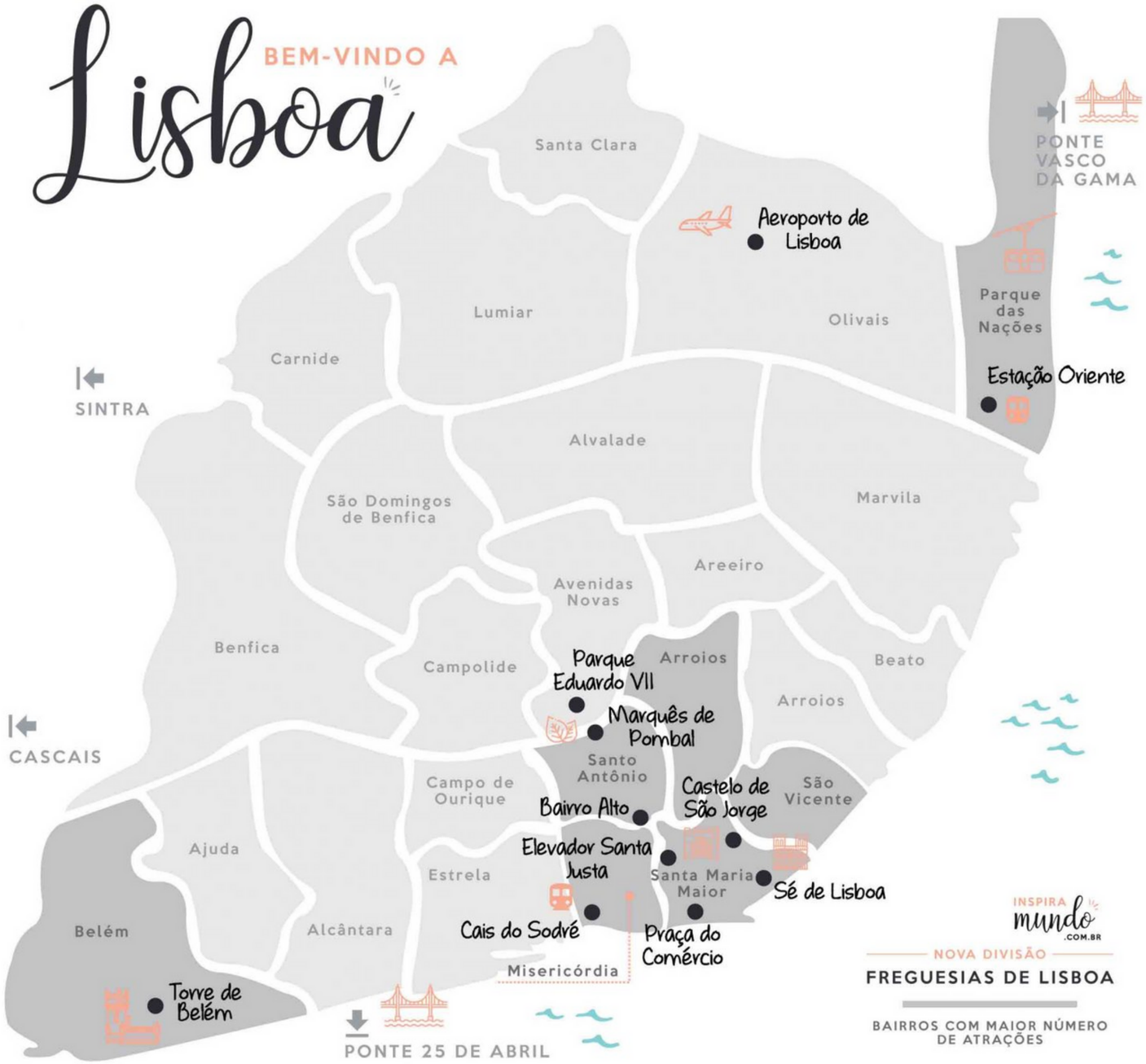
RITA PEREIRA

PAOLA PELAEZ

What We'll Discuss

TOPIC OUTLINE

Business Question
Data
Data Preparation and Cleaning
Main Insights
Main challenges & Learnings

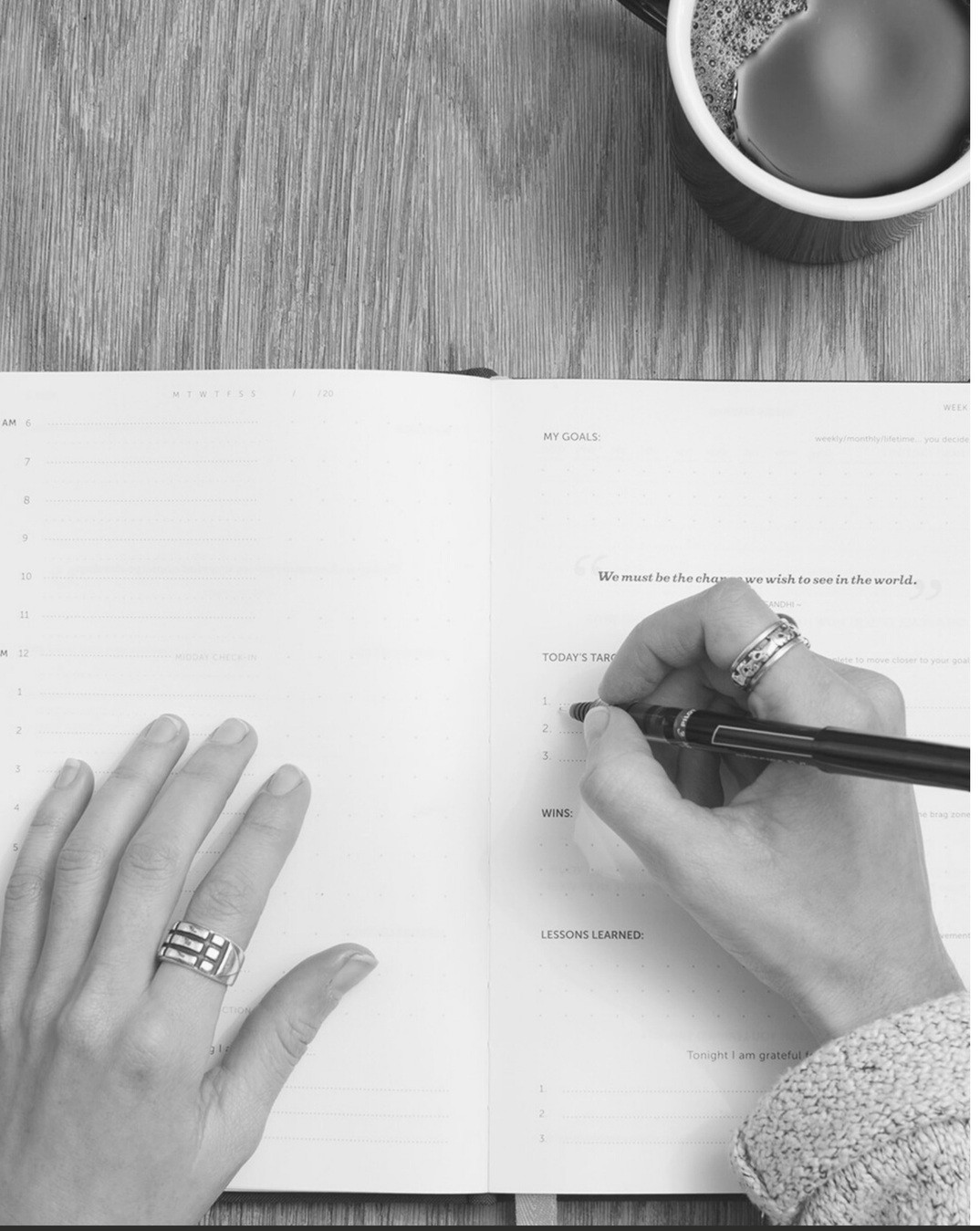


24 Freguesias

53 Bairros

6M TOURISTS*

*Turismo de Portugal 2020



Business Question

Which is the neighbourhood that gives you the best value for money in an Airbnb in Lisbon?

- Which are the most expensive and the cheapest neighborhoods?
- Which is the neighbourhood with best rated Airbnb's in Lisbon?
- Which is the best neighborhood for each kind of traveler: 3 personas (Family, Business and Backpacker) ?

Data

SOURCE: Inside Airbnb
Adding data to the debate

The data covers all Lisbon listing details, customer reviews and associated geolocation information collected on 28 January, 2020 and is published in a form of csv file.
listing_details.csv - Detailed Listings data for Lisbon
(25.002 rows, 106 columns)

LIMITATIONS:
Availability of data to answer question's we would also liked to answer.

Data Preparation and Cleaning

CLEANING

We reduced the number of columns that didn't carry information that could be used to answer our questions:

- Pictures URL's (for hosts, guests, listings)
- Textual descriptions already extracted as continous variables in other columns (space, summary, description)
- Columns that require a lot of preprocessing to turn into useful a feature (transit, access, interaction, house rules)
- Colums that contain no values (host_acceptance_rate)
- Columns that didn't provide added information (city, market, country, country code)
- Colmuns where most of the values were missing or contained only one value (square feet,weekly price, monthly price)

```
#Drop Columns
columns_to_remove = ['neighbourhood',"scrape_id",'listing_url',"last_scraped","experiences_offered","notes","transit","access","
                     "thumbnail_url","house_rules","medium_url","picture_url","xl_picture_url","host_name","host_about",
                     "host_acceptance_rate","host_url","host_thumbnail_url","host_picture_url","host_neighbourhood","host_verifi
                     "state","market","smart_location","country",'country_code',"minimum_minimum_nights","maximum_minimum_nights'
                     "minimum_maximum_nights",'maximum_maximum_nights',"minimum_nights_avg_ntm","maximum_nights_avg_ntm","calend
                     "jurisdiction_names", "street", "calendar_updated", "has_availability", "is_location_exact", "city", "zipco
                     "is_business_travel_ready", "weekly_price", "monthly_price",'maximum_nights', 'availability_30',
                     'availability_60','availability_90','availability_365','square_feet', 'latitude','longitude',
                     "host_id","host_since", "host_location", "host_response_time",'host_response_rate','host_listings_count',
                     'host_total_listings_count','host_has_profile_pic','host_identity_verified','summary','space',
                     'review_scores_accuracy','review_scores_cleanliness','review_scores_checkin','review_scores_communication',
                     'review_scores_location','review_scores_value', 'cleaning_fee', 'security_deposit','extra_people',
                     'requires_license','license', 'calculated_host_listings_count_entire_homes','calculated_host_listings_count_
                     'calculated_host_listings_count_shared_rooms']

data = listings_detailed.drop(columns_to_remove, axis =1)
```

Data Cleaning

CHECKING FOR MISSING VALUES

```
#we have decided to exclude listings with no first nor last review
#the below listings don't have enough information for our quality analysis

null_displ = data[(data['last_review'].isnull()==True)|(data['first_review'].isnull()==True)]
null_displ['review_scores_rating']
```

CHANGING DATA TYPES

```
#let's change price type from object to float
# but first we neeed to remove special characters --> $ symbol
data['price'] = data['price'].str.replace('$', '')
data['price'] = data['price'].str.replace(',', '')

data['price'] = data.price.astype(float)

data['price'].dtypes #sucess!
```

RENAMING COLUMNS

```
#renaming columns so that it's more intuitive
data = data.rename(columns={'neighbourhood_cleansed':'neighbourhood',
                            'neighbourhood_group_cleansed':'city'})
```

FINAL DATAFRAME:
15.733 ROWS
28 COLUMNS

Main Insights

LISBON AIRBNB

General Insights

NEIGHBOURHOOD WITH HIGHER NUMBER OF LISTINGS

SANTA MARIA MAIOR

NEIGHBOURHOOD WITH LOWER NUMBER OF LISTINGS

SANTA CLARA

AVERAGE PRICE PER NIGHT OF AN AIRBNB IN LISBON

€80.65

NEIGHBOURHOOD WITH HIGHER NUMBER OF ENTIRE HOUSE/APT

SANTA MARIA MAIOR

OLDEST AIRBNB LISTING IN LISBON

AVENIDAS NOVAS
9.5 YEARS

MAXIMUM PRICE PER NIGHT OF AN AIRBNB IN LISBON

€1.100

Price Analysis

MOST EXPENSIVE

The neighbourhood with the most expensive mean price is 'Parque das Nações'
Highest price variation in Marvila, Lumiar, Santa Maria Maior, Santa Clara e Arroios

neighbourhood	count	mean	std	min	25%	50%	75%	max
Parque das Naes	305.0	100.045902	67.038367	20.0	60.00	86.0	123.00	500.0
Lumiar	127.0	91.543307	155.857931	13.0	35.00	50.0	84.00	1000.0
Misericrdia	2786.0	87.562455	74.084370	12.0	50.00	70.0	100.00	1000.0
Santo Antnio	1292.0	87.094427	65.751336	10.0	50.00	72.5	100.00	1100.0
Santa Maria Maior	3388.0	87.006494	64.718280	9.0	50.00	72.0	100.00	1000.0

CHEAPEST

neighbourhood	count	mean	std	min	25%	50%	75%	max
Beato	81.0	52.679012	26.076247	10.0	35.00	50.0	69.00	137.0
Olivais	176.0	55.301136	37.940633	11.0	30.00	49.0	70.00	385.0
Penha de Frana	521.0	56.763916	33.865968	9.0	32.00	50.0	75.00	250.0
Carnide	40.0	58.175000	45.664825	9.0	35.00	50.0	66.00	300.0
Benfica	57.0	58.543860	29.634723	14.0	40.00	55.0	70.00	165.0

The neighbourhood with cheapest mean price listings is 'Beato'

Price Analysis

ROOM TYPE-PRICE

Shared room has the lowest price mean

Highest price is in the Hotel Room type

room_type	count	mean	std	min	25%	50%	75%	max
Shared room	170.0	18.735294	8.897539	9.0	13.0	16.0	20.00	60.0
Private room	3170.0	42.717981	38.627502	8.0	25.0	35.0	50.00	1000.0
Entire home/apt	12102.0	89.867708	66.249931	9.0	55.0	75.0	100.00	1100.0
Hotel room	278.0	127.410072	176.232793	13.0	50.0	75.5	138.75	1000.0

Price Analysis

ENTIRE HOME/APT

room_type	neighbourhood	mean	median
Entire home/apt	Ajuda	84.057143	60.0
	Alcntara	79.421525	69.0
	Alvalade	86.930233	70.0
	Areeiro	94.331034	75.0
	Arroios	90.365574	75.0
	Avenidas Novas	109.177083	85.0
	Beato	61.051724	50.5
	Belm	76.784232	65.0
	Benfica	71.764706	62.5
	Campo de Ourique	85.272000	70.0
	Campolide	84.409524	70.0
	Carnide	70.666667	60.0
	Estrela	86.063910	69.0
	Lumiar	99.878049	70.0
	Marvila	83.558140	67.0
	Misericrdia	91.805958	75.0
	Olivais	72.558442	60.0
	Parque das Naes	109.250000	95.0
	Penha de Frana	71.882353	65.0
	Santa Clara	179.600000	90.0
	Santa Maria Maior	90.319283	75.0
	Santo Antnio	99.262564	80.0
	So Domingos de Benfica	86.625000	75.0
	So Vicente	84.545455	69.0

vs.

SHARED ROOM

room_type	neighbourhood	mean	median
Shared room	Ajuda	25.000000	25.0
	Alcntara	17.750000	18.0
	Alvalade	19.000000	16.0
	Areeiro	16.041667	15.0
	Arroios	19.965517	15.0
	Avenidas Novas	15.066667	13.0
	Belm	15.000000	15.0
	Benfica	19.666667	20.0
	Campo de Ourique	14.727273	13.0
	Campolide	9.750000	10.0
	Carnide	25.000000	25.0
	Estrela	29.000000	29.0
	Marvila	18.000000	18.0
	Misericrdia	18.454545	17.0
	Olivais	45.000000	45.0
	Parque das Naes	25.000000	20.0
	Penha de Frana	20.600000	14.0
	Santa Maria Maior	20.933333	15.0
	Santo Antnio	20.461538	20.0
	So Vicente	15.500000	13.5

Customer Ratings

neighbourhood	count	mean	std	min	25%	50%	75%	max
Santa Maria Maior	3388.0	81.801063	86.416969	1.0	15.00	51.0	125.00	486.0
Misericrdia	2786.0	76.561378	84.442589	1.0	14.00	48.0	110.00	683.0
Olivais	176.0	64.943182	113.192565	1.0	5.00	20.0	64.00	866.0
So Vicente	1350.0	63.732593	69.022705	1.0	13.00	41.5	89.00	492.0
Santo Antnio	1292.0	53.137771	67.913735	1.0	8.00	26.0	71.25	536.0

MOST REVIEWED

'Santa Maria Maior' is the most rated neighbourhood. This area includes Baixa de Lisboa, Cais de Sodre and Terreiro do Paço,
These are super touristic areas, therefore confirming our hyphotesis that the most reviewed areas are also the ones tourists tend to choose.

REVIEW SCORE RATING

Sorted by number of reviewes - other insight on the most habitual touristic areas!!!!

neighbourhood	count	mean	std	min	25%	50%	75%	max
Santa Maria Maior	3376.0	92.255628	7.561509	20.0	90.00	94.0	97.0	100.0
Misericrdia	2770.0	91.782671	8.157690	20.0	89.00	94.0	97.0	100.0
Arroios	1958.0	90.956588	9.664513	20.0	88.00	93.0	97.0	100.0
So Vicente	1338.0	92.982810	6.597783	40.0	90.00	94.0	97.0	100.0
Santo Antnio	1283.0	91.855027	8.846544	20.0	89.00	94.0	98.0	100.0

Value for Money

```
# value_money = review_scores_rating/price

data.insert(5, "value_money", data["review_scores_rating"]/data['price'], True)
```

FIRST...

Sorted by mean value for money

neighbourhood	count	mean	std	min	25%	50%	75%	max
Santa Clara	17.0	4.675917	2.915218	0.156667	2.475000	4.700000	5.882353	10.777778
Areeiro	325.0	2.355607	1.639764	0.316667	1.276316	1.904762	3.000000	8.909091
Olivais	174.0	2.305395	1.382242	0.259740	1.346154	1.960000	2.890517	8.333333
Beato	80.0	2.239341	1.345456	0.640000	1.389312	1.760482	2.705462	8.000000
Penha de Frana	517.0	2.215976	1.404766	0.200000	1.250000	1.800000	2.800000	10.444444

neighbourhood	count	mean	std	min	25%	50%	75%	max
Areeiro	325.0	2.355607	1.639764	0.316667	1.276316	1.904762	3.000000	8.909091
Penha de Frana	517.0	2.215976	1.404766	0.200000	1.250000	1.800000	2.800000	10.444444
Alvalade	223.0	2.188012	1.443476	0.208889	1.208333	1.724138	2.757143	8.363636
Arroios	1958.0	1.891874	1.241593	0.080000	1.063645	1.549194	2.352206	8.700000
Campo de Ourique	320.0	1.870112	1.440100	0.122500	1.010420	1.512571	2.090104	12.500000

THEN... >200

We explored the best value for money, among the neighbourhoods with most listings

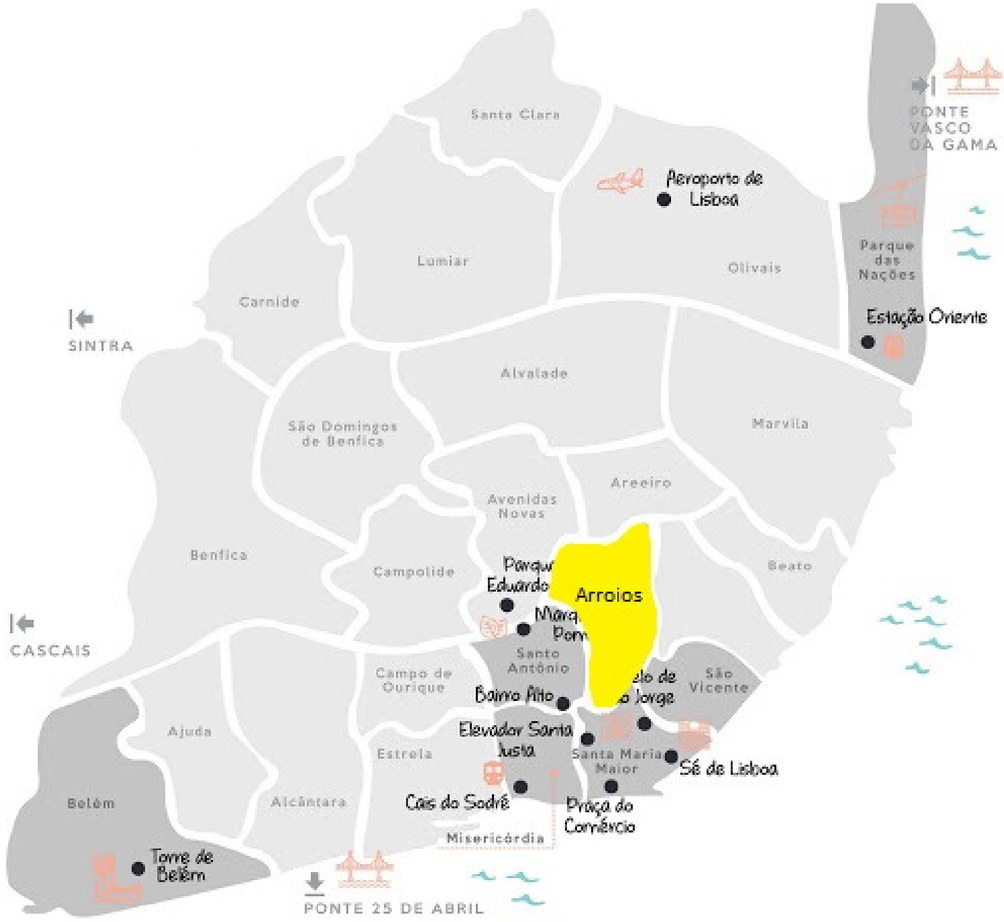
Value for Money

FINALLY... >1000

Among the neighbourhoods with more than 1000 listings,'Arroios'holds best value for money

This is a "new", "hipster", growing area in Lisbon incuding sections like Anjos, Martim Moniz!!

neighbourhood	count	mean	std	min	25%	50%	75%	max
Arroios	1958.0	1.891874	1.241593	0.080000	1.063645	1.549194	2.352206	8.700000
So Vicente	1338.0	1.608294	0.881334	0.103158	1.078652	1.450000	1.900000	8.000000
Santo Antnio	1283.0	1.497327	0.946971	0.077273	0.900000	1.266667	1.820000	7.076923
Santa Maria Maior	3376.0	1.438696	0.867480	0.040000	0.910000	1.280000	1.759444	10.333333
Misericrdia	2770.0	1.427229	0.796267	0.070000	0.928752	1.277047	1.745455	7.384615



Personas

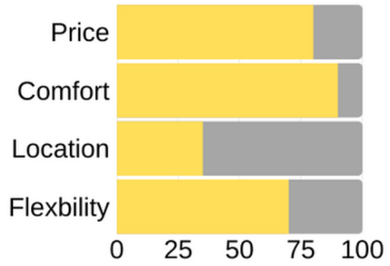
Business

Mike is a regional director who travels 2 to 4 times a month to work.

Goals:
Booking next travel within seven days.

Frustrations:
Too much time spent booking, he is busy!

Motivations:



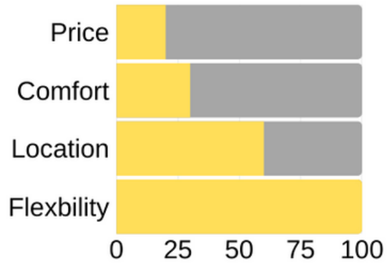
Solo Backpacker

Daniel is backpacking around the world. He wants to collect experiences.

Goals:
Find well located and low budget places.

Frustrations:
Low flexibility to change his plans.

Motivations:



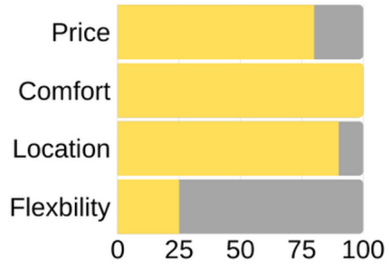
Family

Griswold family travels twice a year and plans their holidays in advance.

Goals:
Find whole and comfortable places.

Frustrations:
Be uncomfortable with the place or location.

Motivations:



Flexibility Analysis

Flexbility

```
# Select by cancellation policy
flexible = data[data ['cancellation_policy'] == 'flexible']  # Business/ Backpacker
moderate = data[data['cancellation_policy'] == 'moderate']  # Family
len(flexible)
len(moderate)
# print(type(flexible))
# flexible.head()
```

```
5792
```

```
# Filter by instant bookable - only flexible
instant_bookable = flexible[flexible['instant_bookable'] == 't']  # Business/ Backpacker
len(instant_bookable)
```

```
1688
```

```
# Filter by require_picture
no_picture = instant_bookable[instant_bookable['require_guest_profile_picture'] == 'f']   # Business/ Backpacker
len(no_picture) #1688

# Filter by require phone verification
the_most_flexible = no_picture[no_picture['require_guest_phone_verification'] == 'f']   # Business/ Backpacker
len(the_most_flexible)
```

```
1681
```

Comfort

```
# Filter by room_type
# Business
business_room_type = the_most_flexible[the_most_flexible['room_type'] == 'Entire home/apt']
len(business_room_type)

# Back
backpacker_room_type = the_most_flexible[the_most_flexible['room_type'] == 'Shared room']
len(backpacker_room_type)

# Family
family_room_type = moderate[moderate['room_type'] == 'Entire home/apt']
len(family_room_type)
```

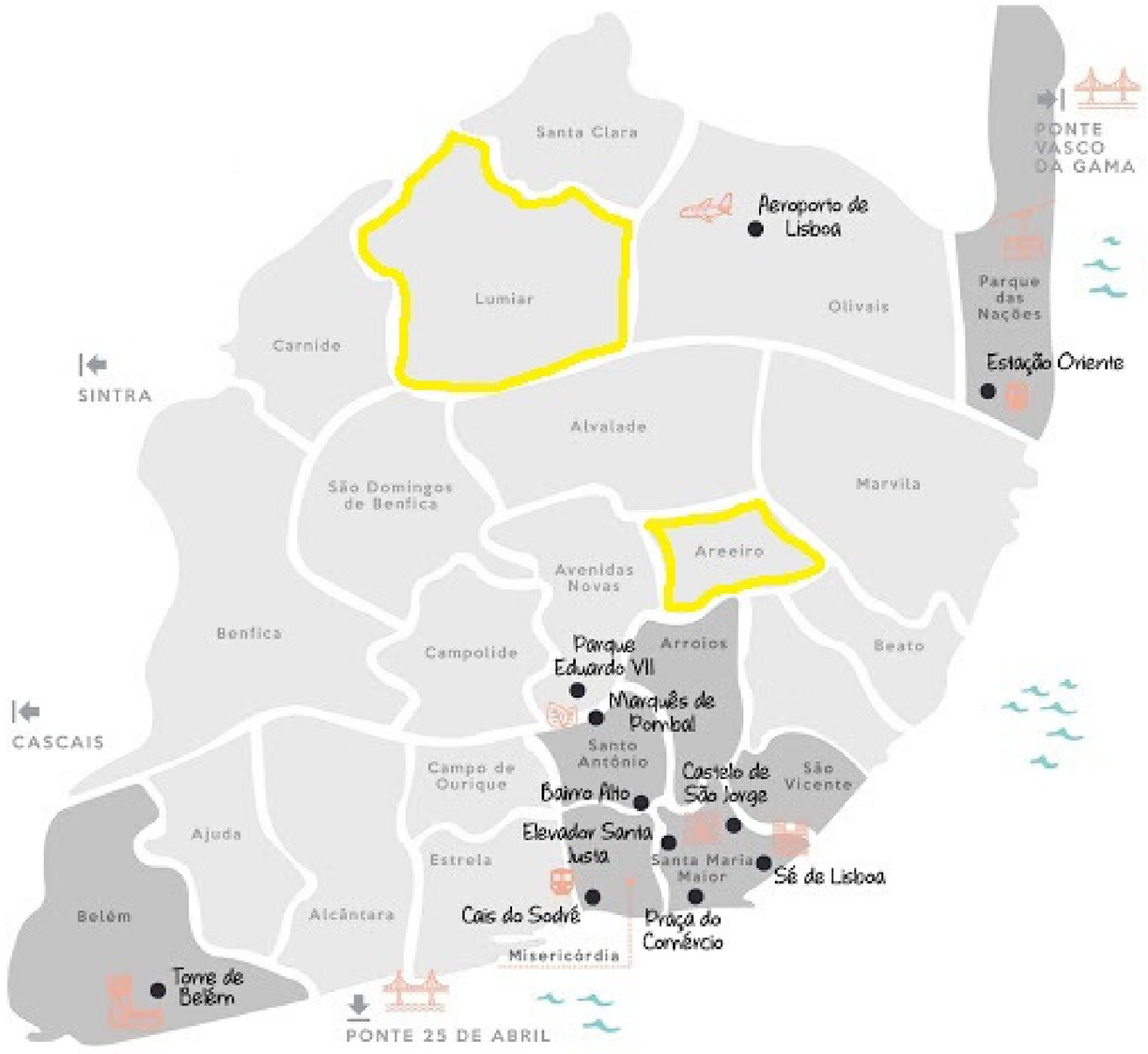
Business

ASCENDING=TRUE

neighbourhood	mean	median
Areeiro	0.928850	0.844000
Ajuda	1.091681	1.058235
Parque das Naes	1.100014	1.083333
So Domingos de Benfica	1.111198	1.242857
Avenidas Novas	1.111842	1.087500

ASCENDING = FALSE

neighbourhood	mean	median
Lumiar	1.919412	1.788372
Marvila	1.675062	1.462687
Beato	1.535826	1.293333
Belm	1.515370	1.470588
Benfica	1.508475	1.508475



Business

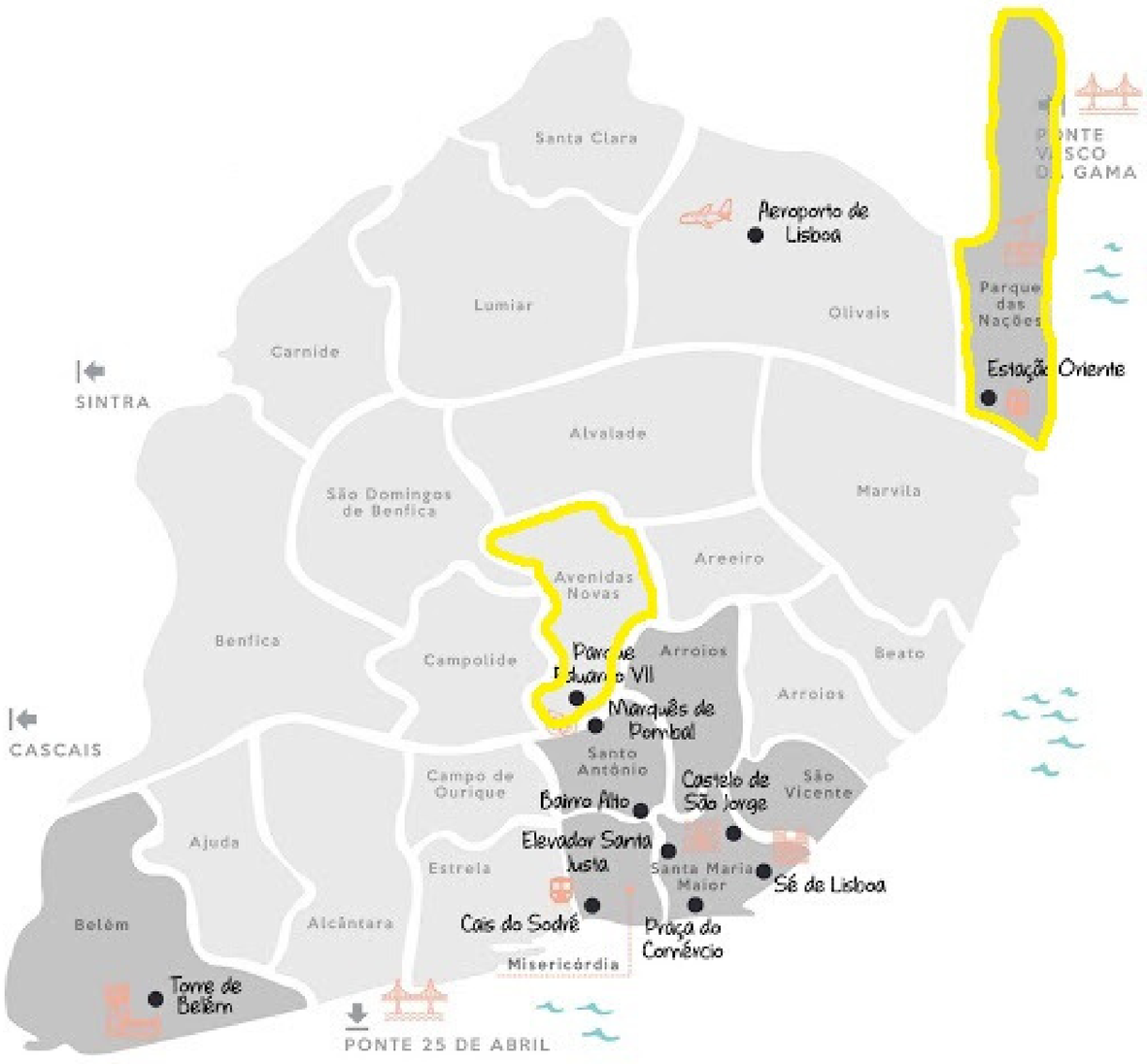
Family

ASCENDING=TRUE

neighbourhood	mean	median
Parque das Naes	1.132051	1.122042
Avenidas Novas	1.174909	1.119048
Santo Antnio	1.211277	1.159420
Misericrdia	1.268604	1.213333
Areeiro	1.310605	1.357143

ASCENDING=FALSE

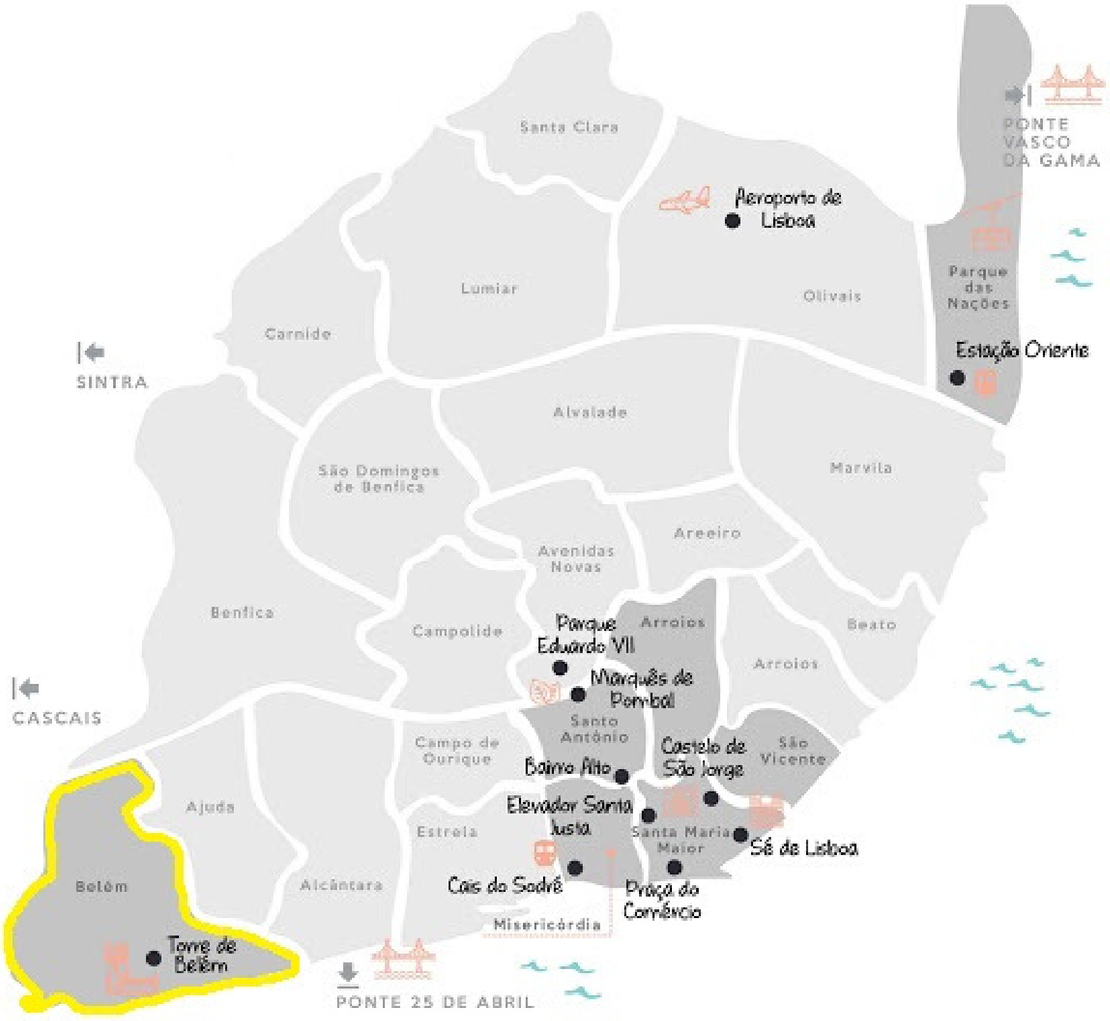
neighbourhood	mean	median
Beato	1.852101	1.720000
Marvila	1.674909	1.627119
Carnide	1.661334	1.417837
Ajuda	1.591197	1.583333
Alcntara	1.566745	1.461538



Family

Solo Backpacker

neighbourhood	mean	median
Belm	6.666667	6.666667
Alcntara	5.555556	5.555556
Penha de Frana	5.555556	5.555556
Santa Maria Maior	5.061905	5.800000
Areeiro	4.926316	4.926316



Solo Backpacker

Main Challenges & Learnings

PUBLIC DATASETS QUALITY

The quality of public datasets is not the best (outdated) or data is not even available.

BUSINESS QUESTION DEFINITION

Finding balance between a good business question and the data available to perform a good analysis.

DATA CLEANING & EXPLORATION

Understanding the data and performing cleaning and manipulation to be able to answer our business question.

THANK YOU!

LET US KNOW IF YOU HAVE QUESTIONS